Approximation in law of Markov processes by non-linear regressions: analytic background and stochastic application

Alex Kulik

Wroclaw University of Science and Technology

Singular diffusions: analytic and stochastic approaches, I

Universität Potsdam, 01.04.19 – 03.04.19

Outline: what are the objects?

- **Diffusions:**

$$dX_t = a(X_t)dt + \sigma(X_t)dW_t$$

- Solution to Lévy-driven SDEs:

$$dX_t = a(X_t)dt + \sigma(X_{t-})dZ_t$$

- **Lévy-type processes:** Markov processes with generator

$$Lf(x) = a(x) \cdot \nabla f(x) + \frac{1}{2} b(x) \cdot \nabla^2 f(x) + \int_{\mathbb{R}} \Big(f(x+z) - f(x) - z \cdot \nabla f(x) 1_{|z|<1} \Big) \mu(x; dz), \quad f \in C^2_\infty \tag{1}$$

Heuristically, a Brownian motion/Lévy process with state dependent characteristics.

One famous example: *Millenial climate changes*.

Ice core data, concentration of CO_2 --→ annual temperature
Available data: 8 000 000 years with step 100, sample of 80 000 points, 30 drastic changes --→ out of range of Gaussian deviations

Ditlevsen 1999: model based on an SDE driven by α-stable noise.

It would be be physically more realistic to have **all** the parameters of the noise state-dependent; e.g. the skewness parameter should be positive in the *cold glacial periods* and negative in the *warmer interstadials* (communicated by I.Pavlyukevich).

Outline: what are the general objectives?

I. 'Mathematical' specification of a 'physically defined' object: existence/uniqueness type results

II. Local properties of the law: existence, bounds, regularity etc of the transition probability density $p_t(x, y)$,

$$P_t(x, dy) = p_t(x, y)dy$$

III. Approximation of *implicit* $p_t(x, y)$ by an *explicit* kernel $p_t^{approx}(x, y)$ with a 'controllable' accuracy

Example: a *Conditionally Gaussian/Euler-Maruyama* approximation of the heat kernel of a diffusion,

$$\begin{aligned} p_t^{EM}(x, y) = (2\pi t)^{-d/2} \Big(\det b(x) \Big)^{1/2} \\ \times \exp\left(-\frac{1}{2t} \Big(b(x)^{-1}(y - x - a(x)t), y - x - a(x)t \Big) \right) \end{aligned} \tag{2}$$

Statistics, I: QMLE

Let 'coefficients' of X depend on $\theta \in \Theta$ and X be observed at discrete time moments $h, 2h, \ldots, nh$.
The *likelikhood function*

$$L_n\left(\theta; x_1, \cdots, x_n\right) = \prod_{k=1}^{n} p_h(\theta; x_{k-1}, x_k)$$

is highly implicit, hence the 'usual' efficient statistical methods such as Maximum Likelikhood Estimation (MLE) are not (easily) applicable.

Prakasa Rao, 1982: replace $p_h(\theta; x, y)$ by $p_h^{EM}(\theta; x, y)$ and maximize the new *Quasi*-likelikhood function. Requires (statistic) **stability condition**

$$nh_n^2 \to 0$$

Pedersen, 1994, Kessler, 1995: better approximations to $p_t(x, y)$ lead to less restrictive stability conditions

The core of the method: to design an approximation $p_t^{approx}(x, y)$ in such a way that the sampled log-derivative

$$\nabla_\theta p_t^{approx}(\theta; X_{t_{k-1}}, X_{t_k}), k = 1, \ldots, n$$

is a 'quasi martingale difference'.

Statistic, II: LA(M)N property

Define *sampled likelihood ratio function*:

$$Z_n(\theta_0, \theta; x_1, \cdots, x_n) = \frac{L_n\left(\theta; X_{t_{1,n}}, \cdots, X_{t_{k,n}}\right)}{L_n\left(\theta_0; X_{t_{1,n}}, \cdots, X_{t_{k,n}}\right)}.$$

The *Local Asymptotic Normality* property holds true at $\theta \in \Theta$ with the matrix rate $r(n)$ and the asymptotic covariance matrix $\Sigma(\theta)$, if for any $u \in \mathbb{R}^d$ if

$$Z_n(\theta, \theta + r(n)u) = \exp\left\{u^\top \Gamma_n(\theta) - \frac{1}{2}u^\top \Sigma(\theta) u + \Psi_n(u, \theta_0)\right\}, \tag{3}$$

$$\Gamma_n(\theta) \Rightarrow_{P^\theta} \mathcal{N}(0, \Sigma(\theta)) \tag{4}$$

$$\Psi_n(u, \theta_0) \to 0, \tag{5}$$

in P^θ-probability. Vaguely,

$$\mathbf{I}_n(\theta)^{-1} = r(n)\Sigma(\theta).$$

LAMN property: $\Sigma(\theta)$ is random, the limit of Γ_n is conditionally Gaussian with the covariance $\Sigma(\theta)$.

Le Cam, 1960: definition and application to construction of estimators

Hajek, 1972: the minimax theorem, an asymptotic version of the Cramer-Rao efficiency bound for **biased** estimators

Requires *relative* bounds rather than absolute ones:

$$\frac{p_t^{approx}(x,y) - p_t(x,y)}{p_t(x,y)}$$

A crucial problem; just think of applying Aronson estimates

$$\frac{\text{upper}}{\text{lower}} \approx \frac{t^{-d/2}e^{-c_1(y-x)^2/t}}{t^{-d/2}e^{-c_2(y-x)^2/t}} = e^{C(y-x)^2/t}$$

with possibly large $C = c_2 - c_1$; not integrable.

One possible solution: Malliavin Calculus approach

Diffusions: Gobet, 2001, 2002

Lévy-driven SDEs:

Corcuera, Kohatsu-Higa
Ivanenko, K., Masuda
Clement, Gloter, Nguyen

The main tool: integral representation of the sensitivity:

$$\nabla_\theta p_t(x,y) = p_t(x,y) E^t_{x,y} \Xi$$

with a certain 'weight' Ξ.
To be discussed in the talk of D.Ivanenko
Hidden limitations:

- requires specific structure of the noise;
- requires SDE representation;
- implicit, and thus hardly can be used for estimation purposes.

Analytic description of $p_t(x,y)$: the parametrix method

Backward Kolmogorov equation:

$$(\partial_t - L_x)p_t(x,y) = 0. \tag{6}$$

'Choose wisely' $p_t^0(x,y)$ and take

$$\Phi_t(x,y) = -(\partial_t - L_x)p_t^0(x,y).$$

Then (6) is transformed to *integral* equation

$$p_t(x,y) = p_t^0(x,y) + (p \circledast \Phi)_t(x,y), \tag{7}$$

$$(p \circledast \Phi)_t(x,y) = \int_0^t \int_{\mathbb{R}^d} p_{t-s}(x,z)\Phi_s(z,y)\,dz,$$

which can be solved as

$$p_t(x,y) = p_t^0(x,y) + (p^0 \circledast \Phi)_t(x,y) + (p^0 \circledast \Phi \circledast \Phi)_t(x,y) + \dots \tag{8}$$

One example: locally α-stable Lévy-type model

A.Kulik, Approximation in law of locally α-stable Lévy-type processes by non-linear regressions, arXiv:1808.06779

- real-valued case $d = 1$;
- no diffusion term $b(x) \equiv 0$;
- *locally α-stable* Lévy kernel: a mixture

$$\mu(x; du) = \mu^{(\alpha)}(x; du) + \nu(x; du).$$

'Principal' α-stable part,

$$\mu^{(\alpha)}(x; du) = \lambda(x) \frac{1 + \rho(x) \operatorname{sgn} u}{|u|^{\alpha+1}} \, du,$$

thanks to state dependent skewness coefficient $\rho(x)$ covers Pavlyukevich's extension of the Dietlevsen model.

The 'nuisance' part $\nu(x; du)$ is allowed to be structured. The main assumption is on the Blumenthal-Getour index, i.e. the intensity of small jumps: for some $\beta < \alpha$,

$$|\nu|(x; \{|z| > r\}) \leq Cr^{-\beta}, \quad r \in (0, 1]. \tag{9}$$

'Microstructural noises' in the spirit of Aït-Sakhalia& Jacod (2006, 2007) are covered.

Conditions

- *compensated drift coefficient*

$$\widetilde{a}(x) = a(x) - 1_{\alpha<1} \int_{|u|\le 1} u\, \mu^{(\alpha)}(x; du) - 1_{\beta<1} \int_{|u|\le 1} u\, \nu(x; du)$$

satisfies

$$|\widetilde{a}(x) - \widetilde{a}(y)| \le C|x-y|^{\eta}, \quad |x-y| \le 1, \tag{10}$$

and the *balance condition* holds:

$$\alpha + \eta > 1 \tag{11}$$

(Tanaka, Tsuchiya, Watanabe 1974: necessity and sufficiency in 1-dim, Kulik 2018: d-dim with isotropic α-stable)

- λ, ρ are ζ-Hölder continuous, for some $0 < \lambda_{\min} < \lambda_{\max}$

$$\lambda_{\min} \le \lambda(x) \le \lambda_{\max}.$$

- The functions $a(x), \nu(x, dz)$ are continuous in x.

Theorem

The martingale problem (L, C_0^2) is well posed in $D(\mathbb{R}^+)$ and, at the same time, the solution X of this martingale problem is the unique Feller process, whose generator A restricted to C_0^∞ coincides with L. This process is strong Feller and possesses a transition probability density $p_t(x, y)$.

Regression approximation

Define *partially compensated* drift coefficient with the truncation level $t^{1/\alpha}$,

$$a_t(x) = a(x) - \int_{t^{1/\alpha}<|u|\leq 1} u\mu(x, du), \tag{12}$$

let[1] a_t be uniformly bounded and uniformly Lipschitz. Define the *regressor function* $\mathfrak{f}_s(x)$ as the solution to the Cauchy problem

$$\frac{d}{dt}\mathfrak{f}_t(x) = a_t(\mathfrak{f}_t(x)), \quad \mathfrak{f}_0(x) = x. \tag{13}$$

Denote by $g^{\lambda,\rho,\upsilon}$ the stable density with intensity parameter λ, skewness parameter ρ, and an external shift υ. Define

$$p_t^{regression}(x, y) = t^{-1/\alpha} g^{(\lambda(x),\rho(x),2\lambda(x)\rho(x))}\left(\frac{y - \mathfrak{f}_t(x)}{t^{1/\alpha}}\right).$$

[1] In this presentation only

Regression approximation

$$p_t(x,y) = p_t^{regression}(x,y) + R_t(x,y)$$

The first summand is the distribution density of

$$\widetilde{X}_t^x = \mathfrak{f}_t(x) + t^{1/\alpha} U^{t,x}, \quad \mathrm{Law}(S^{t,x}) \sim g^{(\lambda_t(x),\rho_t(x),\upsilon_t(x))},$$

a *conditionally stable approximation to* X_t with *regressor* $\mathfrak{f}_t(x)$ and *innovation term* $U^{t,x}$. *Any* regressor would do, which satisfies

$$f_t(x) - \mathfrak{f}_t(x) = o(t^{1/\alpha}).$$

For instance, the Euler regressor $f_t(x) = x + a(x)t$ is OK for $\alpha > 1/2$.

Bounds for the residual term

$$\int_{\mathbb{R}} |R_t(x,y)|\,dy \le Ct^{\delta}, \quad \delta = \min\left[\frac{\alpha+\eta-1}{\alpha}, \frac{\zeta}{\alpha}, 1-\frac{\beta}{\alpha}\right].$$

If $\eta = \zeta = 1, \beta = 0$, the TV-accuracy of approximation $t^{1/\alpha}$.
For diffusions: $t^{1/2}$,

$$|R_t(x,y)| \le Ct^{1/2}t^{-d/2}\exp\{-c\frac{(y-x)^2}{t}\}.$$

Such *kernel estimates* for Lévy-type setting require additional assumptions on the noise. Actually a line of estimates instead of a fixed one can be obtained:

- integral t^{delta}
- uniform-in-x $t^{-1/\alpha+\delta}$
- kernel,

the second and the third requiring more structure of the noise. A by-product of non-locality of the dynamics.

One application: the LAD estimator for unknown drift coefficient

The objective: to estimate the parametric drift coefficient of (SDE)

$$dX_t = b(\theta; X_t)\, dt + \sigma(X_{t-})\, dZ_t^{(\alpha)} + \Sigma(X_{t-})\, dU_t$$

on the basis of a discrete-time sample $(X_{t_{k,n}})_{k=0}^n$, where $t_{k,n} = kh_n$ with the sampling step size $h_n \to 0$ as $n \to \infty$, that is, we consider *high-frequency sampling.*

The model is inspired by Aït-Sakhalia& Jacod:

Drift+ Structured noise part + 'microstructural' noise

The Least Absolute Deviation estimator:

$$\widehat{\theta}_n = \operatorname{argmin}_\theta \mathbb{L}_n(\theta)$$

with

$$\mathbb{L}_n(\theta) := \sum_{k=1}^{n} \left| X_{t_{k,n}} - \mathfrak{f}_t(\theta; X_{t_{k-1,n}}) \right|, \tag{14}$$

where $\mathfrak{f}_t(\theta; x)$ is a (fixed) regressor.
Hiroki Masuda has previously studied the performance of the method for (linear) OU type model

- Analogue of LS estimator with the quardatic loss $(\cdot - \cdot)^2$ replaced by the absolute loss $|\cdot - \cdot|$;
- Has many advantages of LSE (e.g. noise insensitive)
- Is *rate efficient* while LSE is **not** because of heavy tails.

The statement

Conditions:

- $b \in H_\eta$, $\eta + \alpha > 1$, some regularity in θ;
- $\sigma \in H_\zeta$, bounded, uniformly non-degenerate;
- *statistic stability* condition:

$$nh_n^{2\delta} \to 0;$$

- the support of $\mu^U(du)$ is bounded by r and

$$r \sup_x |\Sigma'(x)| < 1.$$

The statement: $\widehat{\theta}_n$ is *consistent* and *asymptotically normal* at the rate

$$r(n) = \sqrt{n} h_n^{1-1/\alpha}.$$

The rate is **optimal** in the sense that the Fischer information of the model

$$\mathbb{I}_n(\theta) \asymp r(n)^{-1}$$

A hint to understand the argument

As for LSEs, one has to study (the limit behavior of)

$$\partial_\theta \mathbb{L}_n(\theta), \quad \partial^2_{\theta\theta} \mathbb{L}_n(\theta)$$

$(|x|)' = \operatorname{sgn}(x)$, bounded, requires L_1 bound for the heat kernel only;

$(|x|)'' = \delta_0(x)$, a finite measure, requires uniform bound (but not a kernel one!)

Improving the approximation

Let X be a diffusion,

$$p_t(x,y) \approx p_t^{EM}(x,y)$$

with accuracy $t^{1/2}$ ($\Leftrightarrow$ the difference allows a kernel estimate of the order $t^{1/2}$), the conditionally Gaussian approximation of the order $1/2$.

The Hermite polynomial approximation of the order 1

$$p_t(x,y) \approx p_t^{EM}(x,y)\left(1 + t^2 \sum_{i,j,k} c_{ijk}(x) H_t^{(i,j,k)}(x,y)\right)$$

with accuracy t,

$$c_{ijk}(x) = \frac{1}{4}\sum_{l=1}^{d} \Big(b_{ij}(x)\Big)'_{x_l} b_{kl}(x).$$

Higher order approximations are also available; current project with D.Ivanenko and A.Kohatsu-Higa

Application to LAMN, I: finite observation horizon

$T_n = T$, $a(\theta; x) = a(x), b(\theta; x) = b(\beta; x)$.

Theorem 1.

Let for some $\gamma \in (0, 1]$ $a \in C_b^{\gamma}, b \in C_b^{1,1+\gamma}$. Then the model satisfies the LAMN property w.r.t. parameter β at any point β_0 with the scalar rate $r(n) = n^{-1/2}$ and the asymptotic variance

$$\Sigma(\beta_0) = \frac{1}{2T} \int_0^T \left(\frac{\partial_\beta b(\beta_0; X_t)}{b(\beta_0; X_t)} \right)^2 dt.$$

Application to LAMN, II: infinite observation horizon

$T_n \to \infty$, $a(\theta;x) = a_0(x) + a_1(\alpha;x), b(\theta;x) = b(\beta;x)$.

Theorem 2

Let $a_0 \in C_b^{\gamma}, a_1 \in C_b^{1,1+\gamma}, b \in C_b^{1,1+\gamma}$. Assume

$$nh_n^{1+\gamma} \to 0,$$

the stability condition, and

$$\Sigma_{T_n}(\theta) := \begin{pmatrix} \frac{1}{T_n}\int_0^{T_n} \frac{(\partial_\alpha a(\alpha;X_t))^2}{b(\beta;X_t)}\,dt & 0 \\ 0 & \frac{1}{2T_n}\int_0^{T_n}\left(\frac{\partial_\beta b(\beta;X_t)}{b(\beta;X_t)}\right)^2 dt \end{pmatrix} \to \Sigma(\theta)$$

in $\mathbf{P}^{\theta_0}$-probability, and for any $\varepsilon > 0$

$$\mathbf{E}^{\theta_0}\left|\Sigma_{T_n}(\theta_0) - \mathbf{E}^{\theta_0}[\Sigma_{T_n}(\theta_0)|\mathcal{F}_{\varepsilon T_n}]\right| \to 0, \quad n \to \infty. \tag{15}$$

Then the model satisfies the LAMN property at the point $\theta_0 = (\alpha_0, \beta_0)$ with the matrix rate

$$r(n) = \frac{1}{\sqrt{n}} \begin{pmatrix} 1/\sqrt{h_n} & 0 \\ 0 & 1 \end{pmatrix}$$

and the asymptotic variance $\Sigma(\theta_0)$.

Depending on the limit behaviour $a(\theta_0; x), x \to \infty$, we may get substantially different asymptotic behavior of X: ergodic, zero-recurrent, and transient. In each of these cases the conditions of part of Theorem 2 can be verified.

The proofs are highly technical, but the backbone argument is easy to explain. Instead of bounds for **ratios** (which are difficult to handle) we use regression approximations with bounds for residual parts which are **individual** (i.e. are estimated w/o taking ratios), but **small** (i.e. the accuracy of approximation is high). The latter circumstance allows us to 'cut off' the residual terms using vague L_1 bounds (instead of more accurate L_2 bounds which are not available because of the lack of martingale structure), and analyze further the well structured and explicit regression part.

Such argument is not sensitive to the structure of the process; Lévy-type processes are visible.

Application to simulation, I: Improved weak approximation schemes for diffusions

Question: numeric procedure to calculate

$$P_T f(x).$$

Standard scheme:

$$P_T f(x) \approx (P^{EM}_{T/N})^N f(x)$$

Talay, Tubaro 1990: approximation rate N^{-1}; N steps with accuracy N^{-2} each.
Improved weak schemes, Klöden, Platen 1992: correction of the Euler-Maruyama scheme by auxiliary random 'seeds' with prescribed moments, which would 'kill' the terms with t^2 in the expansion for

$$P_t f(x) - P^{EM}_t f(x).$$

Bodnarchuk, K. 2018: an alternative improved scheme, based on a polynomial regression.

Application to simulation II: Exact approximation schemes for diffusions

Bally, Kohatsu-Higa, 2015
The parametrix representation of the heat kernel

$$p_t(x,y) = p_t^0(x,y) + (p^0 \circledast \Phi)_t(x,y) + (p^0 \circledast \Phi \circledast \Phi)_t(x,y) + \dots$$

can be written in a 'probabilistic' form

$$p_t(x,y) = e^{-1} E \prod_{k=1}^{N_t} \theta_{\tau_k - \tau_{k-1}}(\widehat{X}^x_{\tau_{k-1},\tau_k}) \delta(\widehat{X}^x_t \in dy).$$

Pro's:

- exact simulation scheme

Contra's

- requires arbitrary large size of products (unless a cutoff is involved, and then the scheme becomes quasi-exact);
- $p_t^0(x,y)$ is not a density, $\widehat{X}_t^x$ is not well defined; requires *forward* parametrix expansion rather than *backward*, which requires more smoothness;
- infinite variance of the weight

$$\prod_{k=1}^{N_t} \theta_{\tau_k - \tau_{k-1}}(\widehat{X}^x_{\tau_{k-1},\tau_k}).$$

Anderson, Kohatsu-Higa, Yuasa, SPA, 2019+: improved (forward) parametrix-based expansion, the weights with finite vairances

Ongoing projects and perspectives

Ongoing:

- Asymptotic expansions of arbitrary order for diffusions, based on Hermite polynomials (with Ivanenko, Kohatsu-Higa)
- Stronger approximations based on parametrix and asymptotic expansions (with Bodnarchuk, Ivanenko, Kohatsu-Higa)
- Statistics for Lévy-type models:
  - LA(M)N property (with Kohatsu-Higa, strong stability condition);
  - LAD-type estimators for 'real world models' (with Masuda, Pavlyukevich).
- Parametrix-based simulation for 'non-flat' Lévy-type models (Marcus type equations, with Bodnarchuk, Pavlyukevich).

Perspectives and challenges:

- QMLE study based on Hermite polynomials expansions (essentially the Aït-Sakhalia's program since mid-2000, open to go)
- Asymptotic expansions for 'heat kernels' of Lévy-type processes; the gains would be visible and quite powerful (statistics and simulation for Lévy-type systems without 'accuracy limits')
- No 'Hermite functions' machinery visible, the 'asymptotic scale' is not clear
- One possible way to proceed: instead of 'refining of backward parametrix representation', some other analytic construction can be used in a more straightforward and thus more effective way. **'Three point parametrix'** is a promising candidate (ongoing, with Bogdan)